Self-Ensembling for Visual Domain Adaptation

French et al., ICLR 2018

Outline

- This paper explores the use of self-ensembling (or teacher-student) model for domain adaptation
  - consistency regularization: minimizing the distance between student and teacher network's predictions
    - student network (weights) is the current state of the model
    - teacher network (weights) is the moving average of all previous states of the model
  - consistency regularization is applied on the unlabeled target data
- Ad-hoc techniques used in this model
  - confidence threshold to filter out teacher network's poor predictions
  - modulating normalization statistics of two domains

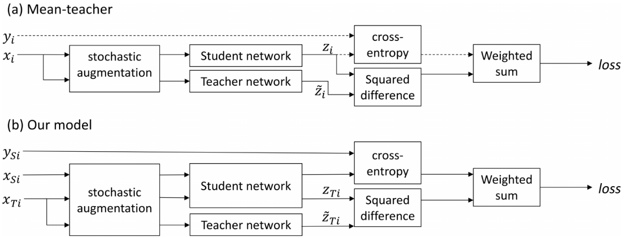
Teacher-Student Model with Consistency Regularization

- Consistency regularization: minimizing the distance between student and teacher network's predictions

$$L_c(x_{\mathrm{ul}}, \theta) = \mathcal{L}(p(y|x_{\mathrm{ul}}; \theta),\ p^+(y|x_{\mathrm{ul}}; \theta^+))$$

- Teacher and student networks
  - weights θ of the student network p is the current state of the main network and is updated via back-propagation
  - weights θ^+ of the teacher network p^+ is an moving average of the past states of the network
    - $\theta^+ = \alpha\theta^+ + (1-\alpha)\theta$, where θ is the current parameters
    - θ^+ is first initialized to be 0

Teacher-Student Model for UDA



- This paper proposes a simple application of mean teacher model to UDA, without use of domain adaptation
  - minimizing the task objective (i.e. cross-entropy) with source labeled data
  - minimizing the consistency regularization with target unlabeled data
  - update student and source accordingly

Ad-hoc Techniques for UDA

- Confidence threshold to discard poor teacher network's predictions
  - for the first few epochs, the teacher network has not accumulated enough past weights and is poor in making predictions
  - to mitigate this problem, the paper sets confidence threshold to select only confident predictions (based on the output softmax probability)
- Modulating normalization statistics of two domains
  - batch norm layers are typically used to standardize the input distribution into network layers that help improve the performance significantly
  - batch norm statistics are computed based on training dataset during training and use that statistics for normalization in inference
  - however, because of the nature of domain adaptation, the statistics of source and target domains are different from one another
  - only statistics for target domain is needed during inference, thus this paper creates separated vectors for these statistics, everything else stays the same

Thank you !

Image Classification Benchmark

	USPS – MNIST	MNIST – USPS	SVHN – MNIST	MNIST – SVHN	CIFAR – STL	STL – CIFAR	Syn Digits – SVHN	Syn Signs – GTSRB
TRAIN ON SOURCE								
SupSrc*	77.55 ±0.8	82.03 ±1.16	66.5 ±1.93	25.44 ±2.8	72.84 ±0.61	51.88 ±1.44	86.86 ±0.86	96.95 ±0.36
SupSrc+TF	77.53 ±4.63	95.39 ±0.93	68.65 ±1.5	24.86 ±3.29	75.2 ±0.28	59.06 ±1.02	87.45 ±0.65	97.3 ±0.16
SupSrc+TFA	91.97 ±2.15	96.25 ±0.54	71.73 ±5.73	28.69 ±1.59	75.18 ±0.76	59.38 ±0.58	87.16 ±0.85	98.02 ±0.20
Specific aug.[b]	–	–	–	61.99 ±3.9	–	–	–	–
RevGrad[a] [1]	74.01	91.11	73.91	35.67	66.12	56.91	91.09	88.65
DCRN [2]	73.67	91.8	81.97	40.05	66.37	58.65	–	–
G2A [3]	90.8	92.5	84.70	36.4	–	–	–	–
ADDA [4]	90.1	89.4	76.00	–	–	–	–	–
ATT [5]	–	–	86.20	52.8	–	–	93.1	96.2
SBADA-GAN [6]	97.60	95.04	76.14	61.08	–	–	–	–
ADA [7]	–	–	97.6	–	–	–	91.86	97.66
OUR RESULTS								
MT+TF	98.07 ±2.82	**98.26** ±0.11	99.18 ±0.12	13.96[c] ±4.41	80.08 ±0.25	18.3 ±9.03	15.94 ±0.0	98.63 ±0.09
MT+CT*	92.35 ±8.61	88.14 ±0.34	93.33 ±5.88	33.87[c] ±4.02	77.53 ±0.11	71.65 ±0.67	96.01 ±0.08	98.53 ±0.15
MT+CT+TF	97.28 ±2.74	98.13 ±0.17	98.64 ±0.42	34.15[c] ±3.56	79.73 ±0.45	**74.24** ±0.46	96.51 ±0.08	98.66 ±0.12
MT+CT+TFA	**99.54** ±0.04	98.23 ±0.13	**99.26** ±0.05	37.49[c] ±2.44	**80.09** ±0.31	69.86 ±1.97	**97.11** ±0.04	**99.37** ±0.09
Specific aug.[b]	–	–	–	**97.0**[c] ±0.06	–	–	–	–
TRAIN ON TARGET								
SupTgt*	99.53 ±0.02	97.29 ±0.2	99.59 ±0.08	95.7 ±0.13	67.75 ±2.23	88.86 ±0.38	95.62 ±0.2	98.49 ±0.32
SupTgt+TF	99.62 ±0.04	97.65 ±0.17	99.61 ±0.04	96.19 ±0.1	70.98 ±0.79	89.83 ±0.39	96.18 ±0.09	98.64 ±0.09
SupTgt+TFA	99.62 ±0.03	97.83 ±0.17	99.59 ±0.06	96.65 ±0.11	70.03 ±1.13	90.44 ±0.38	96.59 ±0.09	99.22 ±0.22
Specific aug.[b]	–	–	–	97.16 ±0.05	–	–	–	–